实验部分：

1. 改变数据分布参数，以实现某些情况下效果提升至20%
2. 进行多分布混合训练
3. 进行纯bandit算法解决问题的训练
4. 在完成上述1,2点后进行bandit与DQN的结合
5. 对比之前RQL算法结果
6. 处理滴滴数据提取特征分布

理论推导部分：

1. 证明固定分布下greedy算法的期望和稳定性。
2. 证明固定模型下opt算法的理论上界
3. 在固定模型和策略下的bandit的regret（先证连续臂只有一个极值点）
4. 证明在分布不变的情况下用马尔科夫模型与opt的比值。从而体现出该方法的适用分布
5. 证明动态分布下DQN结合bandit的期望。